Dependency Distance Motifs in 21 Indo- European Languages

Jing, Yingqi ; Liu, Haitao

Yingqi Jing, Haitao Liu

Dependency Distance Motifs in 21 Indo-European Languages

Abstract: This paper applies the notion of linguistic motif to investigating the linear arrangement of dependency distance (DD) in Indo-European and its implicational meanings in language typology. A series of DD-motifs operating in a decreasing, increasing or equal magnitude are introduced. We first describe the frequency distribution of DD-motifs, and observe a preference for decreasing DD-motifs in human languages. Moreover, we further investigate the role of DD-motifs in controlling the syntactic complexity. The results show that serializing DD values in the same order of magnitude can more or less restrict the structural complexity, and it may be a useful method to realize the DD minimization in natural languages. Finally, we explore the value of DD-motifs in language typology. Our classification experiments reveal that adding the harmonic property and DD-motifs into dependency direction can improve the classification results.

Keywords: DD-motifs, Indo-European languages, structural complexity, language typology

1 Introduction

One of essential traits of the speech is its linear nature, and each element is produced in succession (Saussure 1960: 70). To understand the underlying syntactic structure of a linear sentence, a tree diagram is conventionally adopted (Mel'čuk 1988). According to Tesnière (1959: 19 [2015]), speaking a language involves transforming structural order to linear order, whereas understanding a language involves transforming linear order to structural order. The relationship between linear and structural order is a central topic in formal syntax.

Dependency grammar provides a class of formal descriptions of how pairs of words link in sentences via unequal syntactic relations (Mel'čuk 1988, Hudson 2007). With these syntactic dependencies, we can understand a sentence

Yingqi Jing: Department of Comparative Linguistics, University of Zurich, Zürich, Switzerland, yingqi_jing@163.com
Haitao Liu: Department of Linguistics, Zhejiang University, Hangzhou, China, htliu@163.com

structure by attaching each unit onto a tree diagram, or generate an utterance by projecting the syntactic structure on a timeline. So far, a wide range of evidence suggests that the linear distance between syntactically linked units (called "dependency distance", DD) tends to be minimized in human languages due to universal cognitive constraints that limit the language processing and producing difficulty (Ferrer-i-Cancho 2004, Liu 2008, Gildea and Temperley 2010, Futrell et al. 2015). Many quantitative researches attribute the DD minimization either to the high proportion of adjacent dependencies, or to the rarity of dependency crossing by comparing the natural languages with random languages (Liu 2008, Ferrer-i-Cancho 2013, 2014). Here we think that the randomizing process can also change the DD sequence, which may play certain role in controlling the general structure complexity of languages. Moreover, the DD also has a typological meaning, and Jing (2016) found that the DD can contribute to the improvement of classification effects in word order typology. To proceed, the current study attempts to see whether the sequential arrangement of DD values can be embedded into the language classification.

Specifically, we concentrate on investigating the linear placement patterns of DD and their implications for language typology. For instance, given the statistical tendency for DD minimization, what kinds of DD sequence can significantly reduce the structural complexity of languages? Will the distribution of DD sequence contribute to the improvement of language classification effects? To answer these questions, the current study applied the notion of motifs to invetigating the sequential arrangement of DD in Indo-European languages. The concept of motifs was first put forward by Köhler and Naumann (2008, 2009, 2010) to refer to "the longest continuous sequence of equal or increasing values representing a quantitative property of a linguistic unit". This notion of linguistic motifs was initially inspired by Boroda's F-motiv in musical pieces (Boroda 1982), and can be defined for any linguistic unit (e.g., word, phrase, clause, etc.) in any order (increasing, decreasing or equal).

Here we introduce a new concept called dependency distance motif (DD-motif) to represent the continuous sequence of DD. In so doing, a number of methodological issues or questions remain to be answered. First, due to the differences between head-initial (HI) and head-final (HF) languages, which DD operation, performing in an increasing, decreasing or equal magitude, best faciliates the quantitative analysis of the DD-motifs? Second, will DD-motifs following the same order of magnitude lessen the structural complexity of a language? Finally, if continuous DD sequences are preferred, will the distribution of DD-motifs provide a better classification result in Indo-European?

The rest of this manuscript will first introduce how we can calculate the DD-motifs in a sentence and estimate their structural complexity. 21 Indo-European language dependency treebanks from HArmonized Multi-LanguagE Dependency Treebank (HamleDT) 2.0 are also introduced in this section (Zeman et al. 2012). Section 3 presents the DD-motif results in Indo-European and discusses its value in language typology. The last section is the conclusion.

2 Methods and Materials

Dependency grammar recognizes words as basic syntactic elements, linked via binary asymmetrical relations (Mel'čuk 1988, Hudson 2007, Liu 2009). In a dependency graph (as in Fig. 1), the directed arcs often pointing from the head word to its modifier indicate their dependency relations and directionality. If the head word precedes its dependent in the word string, a HI structure will be formed. Otherwise, it is HF when the head follows the dependent. Dryer (1992, 1996) have observed a harmonic head ordering in natural languages by describing the co-occurring word order features with the order of verb and object. Jing (2016) found that HI or HF dependencies tend to cluster together in the linear sequence. This harmonic property has become a central regularity in word order typology.

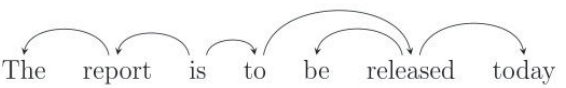

Fig. 1: Dependency graph of a sample sentence

Here we focus on another key concept, dependency distance (DD), which measures the linear distance between syntactically linked words in a sentence. The notion of DD was first put forward by Heringer et al. (1980: 187) and defined by Hudson (1995: 16) as "the distance between words and their parents". To understand or parse a sentence, a word is restored in the memory until it attaches with its head (Gibson 1998). The memory load increases with the length of a dependency (Hudson 2010), and the DD of human languages is generally considered to be minimized due to the constraints of working memory (Ferrer-i-Cancho 2004, Liu 2008, Futrell et al. 2015). Liu (2008) proposed the mean dependency distance (MDD) as a metric for measuring language processing difficulty. This distance-based method for estimating structural complexity is widely used, but it can only measure the distance between syntactically related units

without considering their head orderings or the sequences of DD values. The present research attempts to investigate the linear arrangement of DD and its influence upon the structural complexity of languages. We are seeking to know whether serializing the DD sequences in the same order of magnitude can more or less affects the syntactic complexity of natural languages.

To this end, we apply the definition of motifs from quantitative linguistics to segmenting the DD sequence into continuous and dynamic units. This concept is specifically designed to explore the sequential arrangement of any linguistic elements in a text. A linguistic motif can be defined as "the longest continuous sequence of equal or increasing values representing a quantitative property of a linguistic unit" (Köhler and Naumann 2008, 2009). But we can also generate motifs in other ways, and Köhler and Naumann (2010) suggest that it would be appropriate to go from the last item in the word string to the first unit when analysing a left-branching/head-final language. Here we assume the word order of a sentence going from left to right, and manipulate the order of magnitude by cutting the DD sequence in three different ways: decreasing, increasing, and equal magnitudes. These operations can not only consider the variations between HI and HF languages, but also avoid double counting the same dependencies. Thus, a decreasing DD-motif is a continuous series of reducing DD values, and an increasing or equal DD-motif can be defined accordingly. Moreover, we also segment the DD sequence with respect to the dependency direction (HI or HF). A continuous sequence of DD with the same head direction (HD) is grouped as a consistent HD-motif. It serves as a baseline for comparing with the structural complexity of DD-motifs.

To be more specific, the DD sequences and DD-motifs of the above sample sentence "*The report is to be released today*" can be formed in the following way. We first split the sentence into two parts by the root word, since it has no governor. Second, we can calculate the DD values for each word, and record them sequentially within each domain. Note that we distinguish between the HI and HF dependencies with the symbol -/+. Third, having obtained the DD sequences, we can further generate the DD-motifs according to a decreasing, increasing or equal order. Likewise, we can form the consistent HD-motifs by simply cutting the DD sequences in terms of the dependency direction.

DD sequences:	(+1, +1) (-1, +1, -2, -1)	(1)
Decreasing DD-motifs:	(+1) (+1) (-1) (+1) **(-2, -1)**	
Increasing DD-motifs:	(+1) (+1) (-1) **(+1, -2)** (-1)	
Equal DD-motifs:	**(+1, +1) (-1, +1)** (-2) (-1)	
Consistent HD-motifs:	**(+1,+1)** (-1) (+1) **(-2, -1)**	

In the end, we can extract 5 decreasing DD-motifs, 5 increasing DD-motifs, 4 equal DD-motifs, and 4 consistent HD-motifs from the sentence in Fig. 1. But the current study only focuses on investigating the distribution and structural complexity of DD-motifs with at least two elements. Therefore, 1 decreasing DD-motif, 1 increasing DD-motif, 2 equal DD-motifs, and 2 consistent HD-motifs meet our requirements. We can even look at the dependency direction for each DD-motif. For example, the decreasing DD-motif **(-2, -1)** is consistently HI, and the increasing DD-motif **(+1, -2)** has a mixed head ordering.[1]

To estimate the structural complexity of the DD-motifs, we adopt the metric of MDD from Liu (2008) and calculate the average value for DD-motifs. This

$$MDD\ of\ motifs = sum(DD\text{-}motifs)/n,\ length(DD\text{-}motif) > 1 \quad (2)$$

procedure can be expressed with formula (2).

In this formula, n represents the total number of elements in DD-motifs, whose length is required to be more than 1 word. With this formula, the MDD of consistent HD-motifs in the sample sentence is (1+1+2+1)/ 4= 1.25.

In order to conduct a comparative research on the distribution of DD-motifs in natural languages, we selected 21 Indo-European language dependency treebanks from HamleDT 2.0. They include: 5 Slavic languages (Bulgarian, Czech, Russian, Slovak, Slovenian), 5 Romance languages (Catalan, Spanish, Italian, Portuguese, Romanian), 5 Germanic languages (Danish, German, English, Dutch, Swedish), 3 Indo-Iranian languages (Bengali, Persian, Hindi), 2 Hellenic languages (Greek, Ancient Greek), and 1 Italic language (Latin). These dependency treebanks (or dependency conversions of other treebanks) have been har-

1 The term "consistent" or "mixed" in this study is used to describe the head direction of a dependency. A consistently decreasing DD-motif denotes that all elements in the motif have the same head ordering, and their DD values are arranged in a decreasing magnitude. A mixed DD-motif refers to that those elements in the motif encompass different dependency directions.

monized to the Prague Dependency Treebank annotation style (Zeman et al. 2012). All punctuation and sentences with less than three words were rejected.

3 Results and Discussion

In this section, we begin our investigation of DD sequences by describing the frequency distribution of varying DD-motifs in our sample languages, and we further compare the MDDs of each type of motifs with that of consistent HD-motifs. This exploration can reveal the linear placement patterns of DD in real texts and their influences upon the structural complexity of human languages. Furthermore, we will probe into the role of DD-motifs in language typology.

3.1 Preference for Decreasing DD-motifs

With these dependency treebanks, we computed the number of varying DD-motifs with either consistent or mixed head ordering. We expect our operations can not only reveal the linear regularities of DD but also confirm the tendency for harmonic head ordering (Vennemann 1974, Hawkins 1994, Dryer 1992). This suggests that we should maximize the amount of DD-motifs with the same dependency direction and minimize the mixed DD-motif frequency.

Tab. 1 reports the frequencies of the DD-motifs in 21 Indo-European languages. We first take a look at the total number of DD-motifs in three different orders: decreasing, increasing and equal. The largest figure among the three types of DD-motifs is in boldface. We can see that 11 languages prefer the decreasing DD-motifs, which is slightly higher than that of increasing DD-motifs (10 languages), and the amount of equal DD-motifs is always the least. It seems that the DD sequences with the same magnitude are not preferred in real texts, which is likely due to the selection for a flatter structure in a sentence (Hawkins 1994, Jing and Liu 2015). Because equal DD-motifs largely occur in a chain of adjacent dependencies, which may in turn bring about deeper syntactic structures.

Next, we conduct a detailed comparison of DD-motifs with either consistent or mixed head ordering between three groups. For numerals, decreasing DD-motifs have the largest number of consistent dependencies in 20 languages, while increasing DD-motifs owns the most mixed dependencies in 19 languages. This suggests that arranging DD values in a decreasing magnitude can maximally satisfy the constraints of harmonic principle, and can have less mixed DD-

motifs than operating in an increasing order. It is worth mentioning that though equal DD-motifs can keep the mixed dependencies at the lowest level, the number of consistent dependencies is always the least among the three groups.

In addition, we compare the number of mixed DD-motifs with those of consistent dependency direction within groups. The larger number is put in the box, and we can find that for decreasing DD-motifs all our sample languages except for Persian are inclined to have a consistent head ordering, whereas a tendency for mixed DD-motifs is observed in 19 languages when serializing the DD values in an increasing order. No significant difference between mixed and consistent head ordering is reported in equal DD-motifs. Thus, based on the previous analysis of the distribution of DD-motifs between groups and within groups, we can conclude that human languages prefer a decreasing DD arrangement, since the decreasing DD-motifs can to the most extent preserve the consistent dependencies and restrict the mixed dependencies.

Intriguingly, a roughly complementary distribution of dependency direction between decreasing and increasing DD-motifs is also revealed in Tab. 1. That means suppose a language tends to have consistently decreasing DD-motifs, it is likely to encompass more mixed DD-motifs when performed in an increasing order. Exceptionally, Bengali and Hindi, two completely HF languages, exhibit a consistent head ordering for DD-motifs in any order, and Persian shows a preference for mixed DD-motifs in any order. Further investigations are needed.

Tab. 1: Frequencies of the DD-motifs in 21 Indo-European languages

Lang.	Decreasing			Increasing			Equal		
	Con	Mix	Total	Con	Mix	Total	Con	Mix	Total
Bulgarian (bg)	**28077**	18344	**46421**	14794	**30828**	45622	11176	12696	23872
Czech (cs)	**196084**	147026	343110	103957	**244511**	**348468**	68275	82653	150928
Russian (ru)	**74843**	58219	**133062**	39860	**91315**	131175	33595	35936	69531
Slovak (sk)	**117658**	87587	**205245**	62820	**140725**	203545	40563	46102	86665
Slovenian (sl)	**4533**	3363	**7896**	2311	**5152**	7463	1190	1602	2792
Catalan (ca)	**72314**	45128	**117442**	37938	**79225**	117163	19380	29861	49241
Spanish (es)	**79353**	48940	128293	39917	**89909**	**129826**	21106	33749	54855
Italian (it)	**12665**	6220	**18885**	6752	**11563**	18315	5581	5101	10682
Portuguese (pt)	**36110**	19740	55850	18491	**39822**	**58313**	10547	17266	27813

Romanian (ro)	**5190**	2492	**7682**	2992	**3091**	6083	3354	1579	4933
Danish (da)	**16141**	8196	24337	9591	**15872**	**25463**	6090	4543	10633
German (de)	**93662**	74215	167877	50478	**130213**	**180691**	28840	30948	59788
English (en)	**79275**	48470	127745	35264	**93529**	**128793**	20279	26556	46835
Dutch (nl)	**30346**	15086	45432	14567	**33873**	**48440**	10628	10020	20648
Swedish (sv)	**30983**	19776	50759	18494	**34837**	**53331**	10265	10097	20362
Bengali (bn)	**1385**	159	**1544**	755	**429**	1184	498	61	559
Persian (fa)	**18110**	**29529**	**47639**	13976	27604	41580	10883	12974	23857
Hindi (hi)	51667	**14200**	65867	**53575**	13987	**67562**	33715	6740	40455
Greek (el)	**10994**	8656	19650	4569	**15889**	**20458**	2391	4396	6787
Ancient Greek (grc)	**38639**	34074	**72713**	26316	**44823**	71139	15389	11943	27332
Latin (la)	**6901**	6472	**13373**	5380	**7180**	12560	2881	2316	5197

3.2 Syntactic Complexity for DD-motifs

The preference for decreasing DD-motifs is evidenced by the biased frequency distribution in various language texts. But another question arises: can the linear arrangement of DD values in the same order of magnitude (DD-motif) reduce the structural complexity of languages?

In formal syntax, a substantial evidence points to the universal constraints of DD minimization in human languages (Ferrer-i-Cancho 2004, Liu 2008, Gildea and Temperley 2010, Futrell et al. 2015). Many previous researches attribute this universal property either to the high proportion of adjacent dependencies, or to the rarity of dependency crossing (Liu 2008, Ferrer-i-Cancho 2013, 2014). These two factors are obviously intertwined with each other, since the shorter DD aligns with lower probability of crossing (Ferrer-i-Cancho and Gómez-Rodríguez 2015). A recent study claims that chunking may play an important role in reducing DD (Lu et al. 2016).

These studies on DD minimization are typically conducted by comparing the structure complexity of human languages with that of random languages. Two algorithms, either randomizing dependency relations or changing word order, are applicable in the literature (Liu and Hu 2008, Futrell et al. 2015). It is revealed that natural languages have significantly shorter DD than random languages. Here we argue that those algorithms can not only randomize the linear distance of a dependency, but also alter the sequential arrangement or rhythmic patterns of DD values. We think that comparing natural languages

with random languages can be revealing in certain scenarios, but this paradigm may also bring about some confounding variables. If the placement pattern of DD values is one factor that contributes to the minimization of DD, how can we simply attribute this property to the effects of projectivity or chunking before the exclusion of DD arrangement? More importantly, it is possible to generate random languages without changing the DD of a sentence. Fig. 2 and Fig. 3 construct two random dependency graphs of the sample sentence with the same DD value, but it is hard to tell why these two dependency analyses are not preferred in natural languages. This suggests that the linear distance of a dependency though tends to be minimized, yet the DD minimization is not optimal in human languages (Gildea and Temperley 2010, Futrell et al. 2015) and is not the only mechanism that shapes the complex structure of human languages.

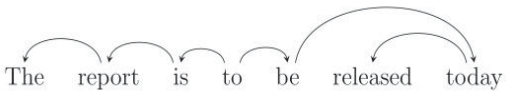

Fig. 2 Dependency graph of the sample sentence with random syntactic relations

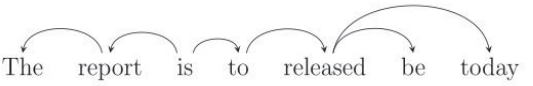

Fig. 3: Dependency graph of the sample sentence with random word order

One observable fact is that the two random languages owns different DD sequences with the natural languages, which leads to our hypothesis that the linear arrangement of DD values may play a certain role in controlling the general structural complexity of languages. Specifically, we wonder whether arranging DD values in the same order of magnitude can to some extent contribute to the DD minimization.

To examine our hypothesis, we evaluate the structural complexity of DD-motifs by measuring their average values of DD (as in formula 2). We first calculate the MDD of consistent HD-motifs, which can be seen as a baseline, since the elements in a consistent HD-motif have the same dependency direction, and their DD values are not placed in the same order of magnitude. After that, we compare it with the MDDs of consistently decreasing, increasing or equal motifs so that the effects of dependency direction can be constrained.

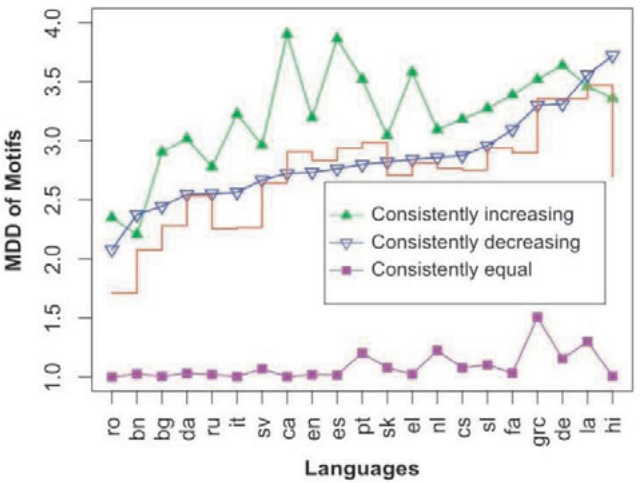

Fig. 4: MDD values for consistently increasing, decreasing or equal motifs

We are expecting that in some languages DD-motifs operating in the same order of magnitude can have MDD values below the baseline.

Fig. 4 shows the average values for consistently increasing, decreasing or equal DD-motifs. The step line represents the baseline value of consistent HD-motifs. We can see that consistently equal DD-motifs in all 21 languages exhibit significantly lower MDD values than the baseline. It is largely caused by the fact that most consistently equal DD-motifs exist in a chain of adjacent dependencies, because their average DD values fluctuate between 1 and 1.5. Moreover, compared with increasing DD sequence, the decreasing DD-motifs have more capability to reduce the structural complexity, since we have observed 6 languages with lower MDD values for consistently decreasing motifs than the baseline, while only 1 language (Latin) shows a lower average value for consistently increasing DD-motifs. This result can be further confirmed by a comparison between the MDDs of varying motifs. 18 languages except for Bengali, Hindi and Latin have lower MDD of consistently decreasing motifs than that of consistently increasing motifs. This finding also echoes the previous distributional preference for decreasing DD-motifs, which we believe is probably the side-effect of the principle of least effort, since human languages favor the sequential arrangement with lower structural complexity (Zipf 1949).

Therefore, we can draw the following conclusion: the DD-motifs can more or less contribute to reducing the structural complexity of natural languages and serializing the DD values in the same order of magnitude may be a useful method to realize the DD minimization.

3.3 Classification Effects for the Skewness of DD-motifs

The above discussion explored the linear arrangement patterns of DD values and their influences upon the general structural complexity of languages. In this part, we are seeking to investigate the role of DD-motifs in language typology.

Modern word order typology is largely concerned with depicting and classifying languages in terms of the placement of certain types of dependencies, say the order of verb and object (Greenberg 1963, Lehmann 1973, 1978, Vennemann 1974). Tesnière (1959: 32) distinguished between HI and HF languages according to the direction of linearization, and Liu (2010) adopted 20 large-scale dependency treebanks to classify languages by their skewness of dependency direction. Jing (2016) probed into the possibility to improve the language classification results by setting the DD as the weight of dependency direction.

These researches have outlined a new dependency treebank-based method to investigate the distribution of dependency direction in human languages, but this approach assumes that each dependency (HI or HF) is independent to each other, disregarding the harmonic property of certain dependencies and the sequential arrangement of DD values. We consider that this model may ignore syntagmatic information in language typology.

To optimize the metric of dependency direction, we proceed by incorporating the property of harmonic head ordering and DD placement patterns into our method. We are trying to see whether the skewness of consistent HD-motifs or the distribution of DD-motifs placing in the same order of magnitude can improve the classification results in Indo-European languages.

We first compare our classification by the skewness of consistent HD-motifs in Fig. 6 with Jing's (2016) classification result in Fig. 5. We can find that two Hellenic languages (Greek, Ancient Greek) are put together on the continuum, though they have shown a huge difference in the dependency direction, which may indicate the word order change from Ancient Greek to modern Greek (Taylor 1994, Liu and Xu 2012). But if we focus on the skewness of continuous HI or HF dependencies, these two languages seem to preserve the similar harmonic head ordering magnitude in the linear sequence. Also, we have witnessed a closer placement of Bulgarian to its genealogical relative, Russian. This poten

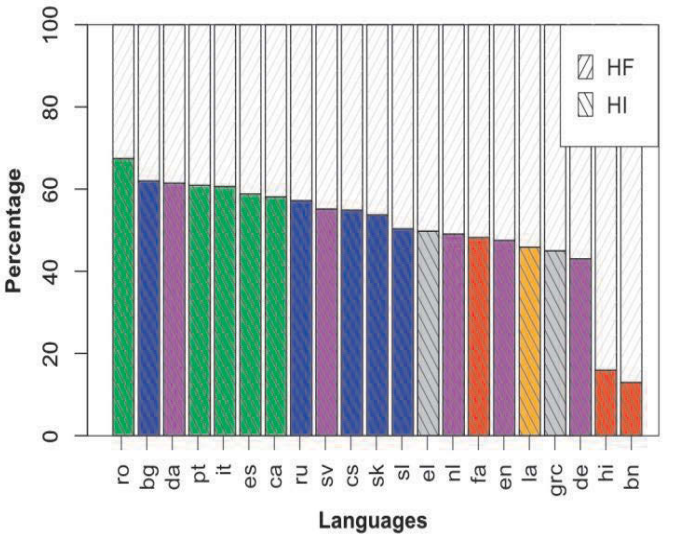

Fig. 5: Language classification result with the skewness of dependency direction (from Jing 2016)

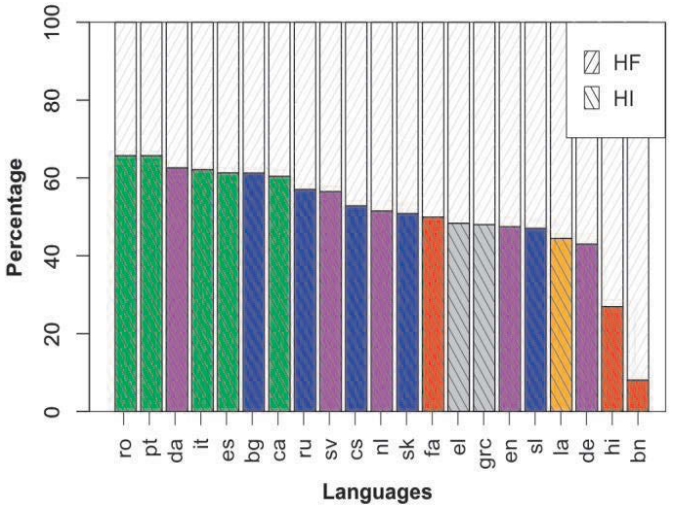

Fig. 6: Language classification with the skewness of consistent HD-motifs

tial shift illustrates that adding the harmonic property into dependency direction can be a useful way to enhance the language classification effects. Notice that Danish is wrongly clustered with Romance languages, which is partly due to its special annotation by setting the determiner as the head of a noun phrase.

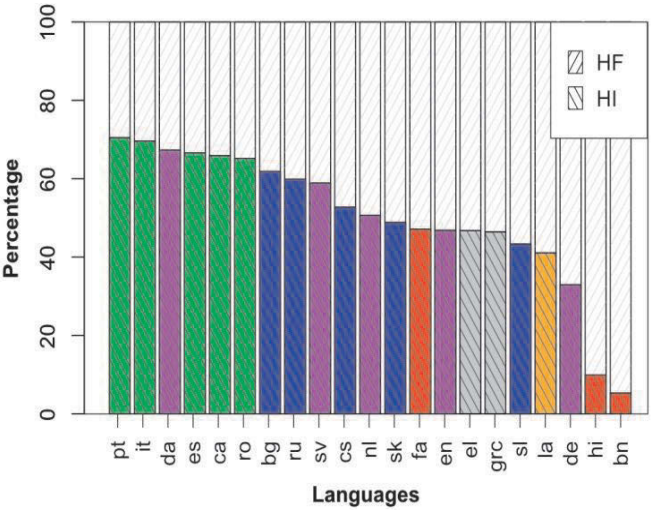

Fig. 7: Language classification with the skewness of consistently decreasing DD-motifs

The normalization process in HamleDT still preserves some of this attribute.

In order to reveal the significance of DD sequence in language typology, we also conducted detailed classifications with the skewness of consistently decreasing, increasing or equal DD-motifs. There are 3! = 6 combinations for the three types of DD-motifs. The optimal classification result is observed for the skewness of decreasing DD-motifs. As shown in Fig. 7, arranging DD sequences in a decreasing order can have a better performance in distinguishing between Romance and Slavic, even though dependency direction often does not work well in classifying languages with free word order (Liu 2010). This improvement we believe can be attributed to the function of consistent DD sequence in the same order of magnitude, since the new metric can give full consideration to the three characters: dependency direction, harmonic property, and sequential arrangement of DD. Interestingly, all these precious metrics work poorly in differentiating between Slavic and Germanic languages. Further researches are necessary to resolve this thorny issue.

4 Conclusion

The current study has conducted a large-scale quantitative analysis of the DD-motifs in Indo-European languages. We concentrate on depicting the frequency distribution and structural complexity of DD-motifs in real texts. We also investigate the value of this unit in language typology. Some major findings are summarized here.

First of all, having described the frequency distribution of three types of DD-motifs (decreasing, increasing or equal) in Indo-European, we find that the segmentation of DD sequence in a decreasing magnitude can maximally preserve the consistent dependencies and effectively control the mixed dependencies. A general preference for decreasing DD-motifs is also observed. Secondly, we further explore the effects of DD-motifs in controlling the MDD of natural languages. By comparing with the MDD of consistent HD-motifs, we have shown that arranging the DD values in a same order of magnitude can more or less ease the language processing difficulty. To be specific, the equal DD-motifs are most capable of lessening the structural complexity of human languages. The decreasing DD-motifs have more ability to restrict the syntactic complexity than the increasing DD-motifs do. Thirdly, DD-motifs are also useful in improving the language classification results. The skewed distribution of consistent HD-motifs can have a better performance in classifying our sample languages than the metric of dependency direction, and an optimal classification result is observed for consistently decreasing DD-motifs.

More quantitative studies applying the DD-motifs to other language families will enrich our findings, or similar researches based on more harmonic language texts will shed some new light on our work.

Acknowledgement

This work is supported by a Ph.D. grant from the China Scholarship Council (201606320224) and the National Social Science Foundation of China under Grant No. 11&ZD188.

References

Boroda, M. (1982). Häuñgkeitsstrukturen Musikalischer Texte. In J. Orlov, M. Boroda, G. Moisei & I. Nadarejšvili (Eds.), *Sprache, Text, Kunst: Quantitative Analysen* (pp. 231-262). Bochum: Studienverlag Dr. N. Brockmeyer.

Dryer, M. S. (1992). The Greenbergian Word Order Correlations. *Language*, 68, 81–138.

Dryer, M. S. (1996). Word Order Typology. In J. Jacobs (Ed.), *Handbook on Syntax, Vol. 2* (pp. 1050–1065). Berlin/New York: Walter de Gruyter Publishing.

Ferrer-i-Cancho, R. (2004). Euclidean Distance between Syntactically Linked Words. *Physical Review E*, 70(5), 056135.

Ferrer-i-Cancho, R. (2013). Hubiness, Length and Crossings and their Relationships in Dependency Trees. *Glottometrics*, 25: 1–21.

Ferrer-i-Cancho, R. (2014). A Stronger Null Hypothesis for Crossing Dependencies. *Europhysics Letters*, 108(5), 58003.

Ferrer-i-Cancho, R., & Gómez-Rodríguez, C. (2015). Crossings as a Side Effect of Dependency Lengths. *http://arxiv.org/abs/1508.06451*

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale Evidence of Dependency Length Minimization in 37 Languages. *PNAS*, 112(33), 10336–10341.

Gibson, E. (1998). Linguistic Complexity: Locality of Syntactic Dependencies. *Cognition*, 68(1), 1–76.

Gildea, D., & Temperley, D. (2010). Do Grammars Minimize Dependency Length? *Cognitive Science*, 34(2), 286–310.

Greenberg, J. H. (1963). Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In J. H. Greenberg (Ed.), *Universals of Human Language* (pp. 73-113). Cambridge, MA: MIT Press.

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Heringer, H.-J., Strecker, B., & Wimmer, R. (1980). *Syntax: Fragen, Lösungen, Alternativen*. München: Fink.

Hudson, R. (1995) Measuring Syntactic Difficulty. Unpublished paper. *http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf*

Hudson, R. (2007). *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.

Hudson, R. (2007). *An Introduction to Word Grammar*. Cambridge: Cambridge University Press.

Jing, Y., & Liu, H. (2015). Mean Hierarchical Distance: Augmenting Mean Dependency Distance. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)* (pp. 161–170). Uppsala, Sweden.

Jing, Y. (2016). Harmonic Head Ordering and Language Classification Optimization (Unpublished Master's thesis). Zhejiang University, P. R. China.

Köhler, R., & Naumann, S. (2008). Quantitative Text Analysis Using L-, F-and T-Segments. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 637–645). Berlin & Heidelberg: Springer.

Köhler, R., & Naumann, S. (2009). A Contribution to Quantitative Studies on the Sentence Level. In R. Köhler (Ed.), *Issues in Quantitative Linguistics* (pp. 34–57). Lüdenscheid: RAM-Verlag.

Köhler, R., & Naumann, S. (2010). A Syntagmatic Approach to Automatic Text Classification. Statistic Properties of F- and L-motifs as Text Characteristics. In P. Grzybek, E. Kelih & J.

Mačutek (Eds.), *Text and Language. Structures, Functions Interrelations, Quantitative Perspectives* (pp. 81–89). Wien: Praesens.
Lehmann, W. (1973). A Structural Principle of Language and its Implications. *Language*, 49, 47–66.
Lehmann, W. (1978). The Great Underlying Ground Plans. In W. Lehmann (Ed.), *Syntactic Typology: Studies in the Phenomenology of Language* (pp. 3-56). Austin: University of Texas Press.
Liu, H. (2008). Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, 9(2), 159–191.
Liu, H., & Hu, F. (2008). What Role does Syntax Play in a Language Network? *Europhysics Letters*, 83(1), 226-234.
Liu, H. (2009). *Dependency Grammar: from Theory to Practice*. Beijing: Science Press.
Liu, H. (2010). Dependency Direction as a Means of Word-order Typology: a Method Based on Dependency Treebanks. *Lingua*, 120(6), 1567–1578.
Liu, H., & Xu, C. (2012). Quantitative Typological Analysis of Romance Languages. *Poznań Studies in Contemporary Linguistics*, 48(4), 597–625.
Lu, Q., Xu, C., & Liu, H. (2016). Can Chunking Reduce Syntactic Complexity of Natural Languages? *Complexity*. doi: 10.100/cplx.21779
Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. Albany: State University Press of New York.
Saussure, F. (1960). *Course in General Linguistics*. London: P. Owen.
Taylor, A. (1994). The Change from SOV to SVO in Ancient Greek. *Language Variation and Change*, 6(1), 1–37.
Tesnière, L. (1959). *Éléments de Syntaxe Structural*. Paris: Klincksieck.
Tesnière, L. (2015). *Elements of Structural Syntax*. Translated by T. Osborne & S. Kahane. Amsterdam: John Benjamins.
Vennemann, T. (1974). Theoretical Word Order Studies: Results and Problems. *Papiere zur Linguistik*, 7, 5–25.
Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., & Hajič, J. (2012). HamleDT: To Parse or not to Parse? In *Proceedings of LREC-2012* (pp. 2735–2741). İstanbul, Turkey.
Zipf, G. (1949). Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Cambridge, Mass: Addison-Wesley Press.

Appendix

Hamle DT 2.0 provides a collection of the 21 Indo-European language treebanks used in our study. They are BulTreeBank/CoNLL 2006, Prague Dependency Treebank, SynTagRus, Slovak Treebank, Slovene Dependency Treebank, AnCora-CA, AnCora-ES, Italian Syntactic-Semantic Treebank/CoNLL 2007, CoNLL 2006 (Floresta Sintá(c)tica), Resurse pentru Gramaticile de Dependenta, CoNLL 2006, TIGER Corpus/CoNLL 2009, Penn Treebank/CoNLL 2007, CoNLL 2006 (Alpino), CoNLL 2006 (Talbanken05), Hyderabad Dependency Treebank/ICON

2010, Persian Dependency Treebank, Hyderabad Dependency Treebank/COLING 2012, Greek Dependency Treebank/CoNLL 2007, Ancient Greek Dependency Treebank, Latin Dependency Treebank. The following documents are related with these treebanks.

Afonso, S., Bick, E., Haber, R., & Santos, D. (2002). Floresta Sintá(c)tica: A Treebank for Portuguese. In *LREC-2002*, Las Palmas, Spain.

Bamman, D., & Crane, G. (2011). The Ancient Greek and Latin Dependency Treebanks. In C. Sporleder, A. Bosch & K. Zervanou (Eds.), *Language Technology for Cultural Heritage, Theory and Applications of Natural Language Processing* (pp. 79-98). Berlin/Heidelberg: Springer.

Bejček, E., Hajičová, E., Hajič J., Jínová, P., Kettnerová V., Kolářová, V., Mikulová M., Mírovský J., Nedoluzhko A., Panevová J., Poláková L., Ševčíková M., Štěpánek J., & Zikánová S. (2013). Prague Dependency Treebank 3.0. *http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3*. Charles University in Prague, ÚFAL, Praha, Czechia.

Boguslavsky, I., Grigorieva, S., Grigoriev, N., Kreidlin, L., & Frid, N. (2000). Dependency Treebank for Russian: Concept, Tools, Types of Information. In *Proceedings of the 18th Conference on Computational Linguistics, Vol. 2* (pp. 987–991). ACL, Morristown, NJ, USA.

Brants, S., Dipper, S., Hansen, S., Lezius, W., & Smith, G. (2002). The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol, Bulgaria.

Călăcean, M. (2008). *Data-driven Dependency Parsing for Romanian*. Uppsala Universitet, Uppsala, Sweden.

Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtsky, Z., & Žele, A. (2006). Towards a Slovene Dependency Treebank. In *Proceedings of LREC-2006* (pp. 1388–1391), ELRA, Genova, Italy.

Husain, S., Mannem, P., Ambati, B. R., & Gadde, P. (2010). The ICON-2010 Tools Contest on Indian Language Dependency Parsing. In *Proceedings of ICON-2010 Tools Contest on Indian Language Dependency Parsing* (pp. 1-8). ICON.

Kromann, M. T., Mikkelsen, L., & Lynge, S. K. (2004). Danish Dependency Treebank. *http://code.google.com/p/copenhagen-dependency-treebank/*. København, Denmark.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational linguistics*, 19(2), 313–330.

Montemagni, S., Barsotti F., Battista M., Calzolari N., Corazzari O., Lenci A., Zampolli A., Fanciulli F., Massetani M., Raffaelli R., Basili R., Pazienza M., Saracino D., Zanzotto F., Mana N., Pianesi F., & Delmonte R. (2003). Building the Italian Syntactic-Semantic Treebank. In A. Abeillé (Ed.), *Building and Using Parsed Corpora* (pp.189-210). Dordrecht: Kluwer.

Nivre, J., Nilsson, J., & Hall, J. (2006). Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In *Proceedings of LREC (Vol. 6)*, Genova, Italy.

Prokopidis, P., Desipri, E., Koutsombogera, M., Papageorgiou, H., & Piperidis, S. (2005). Theoretical and Practical Issues in the Construction of a Greek Dependency Treebank. In *Proceedings of TLT-2005* (pp. 149–160). Barcelona, Spain.

Rasooli, M. S., Moloodi A., Kouhestani M., & Minaei-Bidgoli B. (2011). A Syntactic Valency Lexicon for Persian Verbs: The First Steps towards Persian Dependency Treebank. In *5th Language and Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics* (pp. 227–231). Poznań, Poland.

Rosa, R., Masek, J., Marecek, D., Popel, M., Zeman, D., & Zabokrtský, Z. (2014). HamleDT 2.0: Thirty Dependency Treebanks Stanfordized. In *Proceedings of LREC-2014* (pp. 2334–2341). ELRA, Reykjavík, Iceland.

Simov, K., & Osenova P. (2005). Extending the Annotation of BulTreeBank: Phase 2. In *Proceedings of TLT-2005* (pp. 173–184). Barcelona, Spain.

Šimková, M., & Garabík R. (2006). Синтаксическая разметка в Словацком национальном корпусе. In *Труды международной конференции Корпусная лингвистика – 2006* (pp. 389–394). Sankt-Peterburg: St. Petersburg University Press.

Taulé, M., Antònia M., & Recasens M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of LREC-2008*. ELRA, Marrakech, Morocco.

Van der Beek, L., Bouma, G., Malouf, R., & Van Noord, G. (2002). The Alpino Dependency Treebank. *Language and Computers*, 45(1), 8–22.